Feature Compensation Scheme Based on Parallel Combined Mixture Model

Wooil Kim, Sungjoo Ahn and Hanseok Ko

Dept. of Electronics and Computer Engineering
Korea University, Seoul, Korea
{wikim,sjahn,hsko}@korea.ac.kr

Abstract
This paper proposes an effective feature compensation scheme based on speech model for achieving robust speech recognition. Conventional model-based method requires off-line training with noisy speech database and is not suitable for online adaptation. In the proposed scheme, we can relax the off-line training with noisy speech database by employing the parallel model combination technique for estimation of correction factors. Applying the model combination process over to the mixture model alone as opposed to entire HMM makes the online model combination possible. Exploiting the availability of noise model from off-line sources, we accomplish the online adaptation via MAP(Maximum A Posteriori) estimation. In addition, the online channel estimation procedure is induced within the proposed framework. The representative experimental results indicate that the suggested algorithm is effective in realizing robust speech recognition under the combined adverse conditions of additive background noise and channel distortion.

1. Introduction
The difference between training and operating environment is a significant factor affecting but mostly degrading the performance of speech recognition system. How to make both conditions equal is one of the most essential issues in the development of the actual applications of speech recognition technology and vigorous research efforts have been pursued to realize this goal. As an effort to bring the operating environment closer to the training environment, various techniques have been used at the pre-processing level of speech recognition system. Spectral subtraction, CMN(Cepstral Mean Normalization), model based feature compensation are some of the prominent examples [1][2]. Another approach recently introduced is based on the acoustic speech model transformation. It aims at not removing the noise components but generating the speech model matched to the noisy environment. MAP and MLLR(Maximum Likelihood Linear Regression) adaptation techniques and PMC(Parallel Model Combination) method are included in this category [3][4].

In this paper, we focus on the Gaussian mixture model (GMM) based feature compensation method to render improved recognition under the combined adverse conditions of additive background noise and channel distortion. Employing the model combination, the proposed scheme does not require training with noisy speech database, which is indispensable to the conventional methods. Independent access to the noise model makes its adaptation in the non-speech interval possible. The online channel estimation induced from the proposed algorithm enables the system to realize more robustness to the convolution noise.

The paper is organized as follows. We first review the GMM based feature compensation and identify the relevant issues in Section 2. We then describe the proposed scheme in Section 3. The representative experimental procedures and results are presented and discussed in Section 4. Finally, in Section 5, we make our concluding remarks and discuss future works.

2. Gaussian mixture model based feature compensation: RATZ
The feature compensation method based on speech model was first proposed by Acero and soon afterwards, Moreno designed a data-driven method called RATZ (Multivariate Gaussian Based Cepstral Normalization), which has been motivating similar schemes elsewhere. In RATZ, the statistical transformation of the clean speech’s cepstral distribution under the noisy condition is estimated from the noisy speech database and then noisy input feature vectors are compensated using these statistics [3][5].

Distribution of clean speech cepstrum can be modeled with \( K \) Gaussian mixture as follows.

\[
p(x) = \sum_{i=1}^{K} a_i N_s(\mu_{x,i}, \sum_{x,i})
\]

(1)

It is assumed that noisy environment causes the shift in the means and the compression or expansion of the
covariance matrices of cepstral distributions. Therefore, we can express the distribution of noisy speech as
\[
p(y) = \sum_{k=1}^{K} a_k N_r(\mu_{x,k} + r_k, \Sigma_{x,k} + R_k).
\] (2)

Noisy input feature vectors are then compensated based on the MMSE (Minimum Mean Squared Error) estimator.
\[
\hat{x}_{\text{noise}} = E[x \mid y] = \int x \cdot p(x \mid y) dx
\]
\[
\cong y - \sum_{k=i}^{K} r_k p[k \mid y]
\] (3)

Correction factors \( r_k, R_k \) are obtained through training with the noisy speech database constructed under the identical environment with the testing condition.

RATZ has the advantage of simple and fast computational procedure, however, encounters the following problems when applied to the actual condition.
1) Estimation of correction factors requires training with the noisy speech database.
2) The performance degrades drastically when the training condition does not match to testing one.
3) Online adaptation technique is not appropriate since model updating for RATZ needs sufficient utterances. Although correction factors estimation in RATZ does not require a huge speech corpus as much as HMM training, construction of noisy speech database matched to testing condition is a laborious work. Under the real life condition such as driving in a car, the noisy condition changes along the time. Therefore, the system is required to update the parameters for feature compensation adaptively. However, model estimation in the existing RATZ demands the speech utterances enough to overcome the “sparse-data” problem, thereby making the online adaptation technique not suitable to RATZ.

In this paper, we intend to eliminate the training procedure with the noisy speech database by applying PMC (Parallel Model Combination) in the estimation of correction factors. Since employing PMC means the ability for independent access to noise model, the noise model adaptation is reasonable in the proposed framework. Introducing the channel normalization is expected to make the system more robust, which is not taken into account in the PMC work.

3. Proposed Algorithm
3.1. Parallel Model Combination
In PMC, assuming that a recognition system has the optimal performance when the training and testing condition is identical each other, the clean speech model is transformed to the corrupted speech model matched to the actual noisy environment. To generate the noise-corrupted speech model, the clean speech model and noise model are used independently. PMC is known to exhibit an outstanding advantage in that it does not require additional training procedure with noisy speech database [2].

In log-normal approximation, it is assumed that the addition of two log-normal distributions results in a log-normal form also. The mean and covariance of corrupted speech are computed by Equation (4) under such assumption. In (4), \( \hat{\mu}, \mu, \mu, \hat{\mu} \) refer to the mean vectors of corrupted speech, clean speech and noise respectively and \( \hat{\Sigma}, \Sigma, \hat{\Sigma} \) denote their covariance matrices in log-normal distributions.
\[
\hat{\mu} = g\mu + \hat{\mu}
\]
\[
\hat{\Sigma} = g^2 \Sigma + \hat{\Sigma}
\] (4)

The mean and covariance of linear spectrum that has log-normal distribution are obtained from the mean and covariance of log spectrum by (5).
\[
\mu_l = \exp(\mu'_l + \Sigma'_l / 2)
\]
\[
\Sigma_l = \mu_l \mu_l \left[ \exp(\Sigma'_l) - 1 \right]
\] (5)

Finally, we calculate the mean and covariance of corrupted speech’s log spectrum by the following equations.
\[
\hat{\mu}'_l = \log(\hat{\mu}_l) - \frac{1}{2} \log \left( \frac{\hat{\Sigma}_l}{\hat{\mu}_l} + 1 \right)
\]
\[
\hat{\Sigma}'_l = \log \left( \frac{\hat{\Sigma}_l}{\hat{\mu}_l \hat{\mu}_l} + 1 \right)
\] (6)

3.2. Parallel Combined Mixture Model Based Feature Compensation
As in RATZ, the proposed PCMM (Parallel Combined Mixture Model) based feature compensation method utilizes the statistical distribution of speech feature in the form of cepstrum. PCMM estimates the correction factors from the noisy speech mixture model generated by combining the clean speech model and noise model, whereas RATZ obtains them from the noisy speech database training. Mean and covariance in cepstrum domain can be transformed to those of log-spectrum domain by inverse DCT (Discrete Cosine Transform). Model parameters of noisy cepstrum distribution are generated by model combining process, thus, combination of clean speech model and noise model, and return to the cepstrum domain via DCT transform. Considering the variations in mean and covariance which are assumed in RATZ, we can compute the correction factors as follows.
\[
r = \mu - \hat{\mu},
\]
\[
R = \Sigma - \hat{\Sigma},
\] (7)
Restoration of clean feature vectors is accomplished using the MMSE identical to RATZ.

The proposed PCMM based feature compensation technique has the following advantages:
1) It no longer requires additional training procedure with the noisy speech database.
2) The model combination processing applied to just the Gaussian mixtures as opposed to the entire HMM achieves significant reduction in the computational load and it makes the online model combination possible.
3) Independent availability of noise model makes the noise model adaptation possible. We can update the noise model itself in the non-speech interval and it is free from the “sparse-data” problem compared to the speech model re-estimation.

3.3. Noise Model Adaptation
Noise model adaptation is accomplished in the non-speech interval by MAP estimation. We expect the short length of input signal to be ‘trainable’, since noise is modeled with single Gaussian distribution. MAP adaptation technique has been developed, in which the model parameters are estimated from incomplete data by applying EM (Expectation Maximization) and introducing prior probability of model parameters [4]. Mean of single Gaussian pdf is updated by the following simple equation.

$$\hat{m} = \frac{m + 1/T \sum_{t} x_{t}}{\tau + 1/T}$$  \hspace{1cm} (8)

That is, the mean of initial noise model $m$ and sample mean of non-speech region are weighted by adaptation rate $\tau$ and summed into a new updated noise mean.

3.4. Online Channel Normalization
Since the general PMC is designed for only additive background noise, we have to remove the convolution noise resulting from channel distortion. In the proposed scheme, we regard the speech corruption process as an approximated equation and induce a sequential EM algorithm for channel estimation. Speech corruption process can be modeled in cepstrum domain by following equation.

$$y = x + h + C\log(1 + \exp(C^{-1}(n - x - h)))$$  \hspace{1cm} (9)

If the channel effect $h$ is assumed to be small relatively compared to the additive noise $n$, (9) can be approximated to (10).

$$y = x + h + C\log(1 + \exp(C^{-1}(n - x))) = y_{s} + h$$  \hspace{1cm} (10)

In (10), $y_{s}$ refers the cepstrum vector corrupted by only additive noise and its statistical parameters are estimated through the proposed PCMM procedure, in which noisy speech model is generated via combining clean speech model and noise model. To find $h$ that maximizes the likelihood of $y$, applying EM algorithm leads to the following solution.

$$\hat{h} = \frac{\sum_{t=1}^{T} \sum_{k=1}^{K} P(k | y_{t}, h) \Sigma_{y_{t}}^{-1}(y_{t} - \mu_{y_{t}})}{\sum_{t=1}^{T} \sum_{k=1}^{K} P(k | y_{t}, h) \Sigma_{y_{t}}^{-1}}$$  \hspace{1cm} (11)

where $\mu_{y_{s}}$ and $\Sigma_{y_{s}}$ are mean and covariance of $y_{s}$ respectively and (11) is modified into the sequential form for the real-time implementation.

$$\hat{h}_{t} = \frac{\sum_{t=1}^{T} \sum_{k=1}^{K} P(k | y_{t}, \hat{h}_{t-1}) \Sigma_{y_{t}}^{-1}(y_{t} - \mu_{y_{t}})}{\sum_{t=1}^{T} \sum_{k=1}^{K} P(k | y_{t}, \hat{h}_{t-1}) \Sigma_{y_{t}}^{-1}}$$  \hspace{1cm} (12)

Fig. 1 shows an entire block diagram of the proposed scheme in this paper.

4. Experiments and results
We followed the Aurora2.0 evaluation procedure for the performance verification. Along with all identical conditions suggested in the Aurora2.0 procedure, we used $c_{0}$ instead of log-energy for the convenience of PCMM implementation.

First, we examined the baseline system’s performance to car noise condition of Set A (clean condition training) in Aurora2.0. For the clean speech modeling in RATZ, 128 Gaussian mixture is used and correction factors are estimated from the multi-condition training data matched to each SNR. Aurora2.0 does not contain 0dB and -5dB’s noisy speech data for training, so we could not obtain the results for those cases. Table 1 shows the baseline system’s performance.

Under the identical condition with baseline test, we accomplished the performance evaluation of the proposed scheme. Assuming that a silence interval exists during the training phase of each utterance in training speech for RATZ, we collected noise samples from those regions and estimated noise models.
matched to each SNR condition. All of the noise models are estimated with a single Gaussian distribution. In the cases of 0dB and -5dB, the noise samples had to be extracted from the test speech data. For the comparison, we examined the performance in the following combinations.

1) PCMM: PCMM based feature compensation
2) PCMM-NA: PCMM + noise adaptation
3) PCMM-CN: PCMM + noise adaptation + channel normalization
4) PCMM-NA-CN: PCMM + noise adaptation + channel normalization

As shown in the 2nd column of Table 2, we can see that the proposed feature compensation method is effective under the noisy condition and these figures present its desirable performance comparing to RATZ. The results prove that the model combination used for correction factors estimation effectively represents the noise corruption process. Through the noise model adaptation, we could obtain an improvement over basic PCMM by 1.81% in word accuracy. It says that the MAP adaptation in the silence interval reflects the change of noise at each utterance appropriately. The online channel normalization employed in this paper brought 1.46% gain over basic PCMM. Although the noisy samples do not have channel distortion factors, the proposed algorithm has an effect of normalizing the variations across the speakers. PCMM with both noise adaptation and channel normalization had the best results and it shows the effectiveness of the proposed scheme under additive background noise and convolution noise (e.g. channel distortion).

To observe the effects of adaptation and channel normalization more pronounced, we examined the performance in the channel distorted test samples. Subway (MIRS) of Set C has test samples attenuated in the lower frequency region by a simulated filter like mobile communication environment as well as corrupted by subway train noise additively. A pure subway train noise model is used as an initial noise model, which does not reflect the channel effect. Table 3 presents the obvious improvement in channel distortion condition.

### 5. Conclusions

In this paper, we have proposed a new feature compensation algorithm based on speech model. By applying PMC to correction factors estimation, the proposed scheme does not require training with the noisy speech database. We employed the MAP estimation for noise model adaptation in the non-speech region and introduced online channel estimation within the proposed work. The experiments’ results show that the suggested algorithm is considerably effective in the adverse conditions of additive background noise and channel distortion.

### Table 1. Word accuracy for baseline system to car noise condition in Aurora2.0 (%)

<table>
<thead>
<tr>
<th></th>
<th>Clean</th>
<th>20dB</th>
<th>15dB</th>
<th>10dB</th>
<th>5dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>98.84</td>
<td>96.42</td>
<td>87.62</td>
<td>61.71</td>
<td>26.87</td>
</tr>
<tr>
<td>RATZ</td>
<td>98.84</td>
<td>97.79</td>
<td>96.33</td>
<td>91.95</td>
<td>82.17</td>
</tr>
</tbody>
</table>

### Table 2. Word accuracy for the proposed scheme to car noise condition in Aurora2.0 (%)

<table>
<thead>
<tr>
<th></th>
<th>Clean</th>
<th>20dB</th>
<th>15dB</th>
<th>10dB</th>
<th>5dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>98.84</td>
<td>98.84</td>
<td>98.81</td>
<td>98.66</td>
<td>98.72</td>
</tr>
<tr>
<td>PCMM</td>
<td>96.42</td>
<td>97.61</td>
<td>97.97</td>
<td>98.21</td>
<td>98.30</td>
</tr>
<tr>
<td>PCMM-NA</td>
<td>87.62</td>
<td>96.57</td>
<td>96.75</td>
<td>97.29</td>
<td>97.41</td>
</tr>
<tr>
<td>PCMM-CN</td>
<td>61.71</td>
<td>91.32</td>
<td>92.81</td>
<td>93.59</td>
<td>94.36</td>
</tr>
<tr>
<td>PCMM-NA-CN</td>
<td>57.87</td>
<td>77.51</td>
<td>80.79</td>
<td>79.66</td>
<td>82.22</td>
</tr>
<tr>
<td>0dB</td>
<td>10.38</td>
<td>47.84</td>
<td>51.57</td>
<td>49.39</td>
<td>53.36</td>
</tr>
<tr>
<td>-5dB</td>
<td>8.41</td>
<td>18.73</td>
<td>20.25</td>
<td>18.16</td>
<td>21.12</td>
</tr>
<tr>
<td>Avg</td>
<td>56.60</td>
<td>82.17</td>
<td>83.98</td>
<td>83.63</td>
<td>85.13</td>
</tr>
</tbody>
</table>

### Table 3. Word accuracy for the proposed scheme to Subway(MIRS) condition in Aurora2.0(%)  

<table>
<thead>
<tr>
<th></th>
<th>Clean</th>
<th>20dB</th>
<th>15dB</th>
<th>10dB</th>
<th>5dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>99.08</td>
<td>99.08</td>
<td>99.11</td>
<td>98.93</td>
<td>98.89</td>
</tr>
<tr>
<td>PCMM</td>
<td>95.86</td>
<td>96.44</td>
<td>97.33</td>
<td>97.91</td>
<td>97.91</td>
</tr>
<tr>
<td>PCMM-NA</td>
<td>91.31</td>
<td>93.58</td>
<td>94.84</td>
<td>96.13</td>
<td>96.16</td>
</tr>
<tr>
<td>PCMM-CN</td>
<td>78.29</td>
<td>85.94</td>
<td>89.04</td>
<td>92.23</td>
<td>92.26</td>
</tr>
<tr>
<td>PCMM-NA-CN</td>
<td>43.63</td>
<td>63.62</td>
<td>71.51</td>
<td>84.71</td>
<td>84.92</td>
</tr>
<tr>
<td>0dB</td>
<td>16.70</td>
<td>31.96</td>
<td>38.32</td>
<td>63.89</td>
<td>65.34</td>
</tr>
<tr>
<td>-5dB</td>
<td>8.93</td>
<td>14.34</td>
<td>16.52</td>
<td>31.93</td>
<td>34.76</td>
</tr>
<tr>
<td>Avg</td>
<td>65.16</td>
<td>74.31</td>
<td>78.21</td>
<td>86.97</td>
<td>87.32</td>
</tr>
</tbody>
</table>

### 6. Acknowledgement

This work was supported by Speech Information Technology Research Center.

### 7. References